DOCUMENT RESUME

ED 459 207 TM 033 506

AUTHOR	Hambleton, Ronald K.; Patsula, Liane
TITLE	Adapting Tests for Use in Multiple Languages and Cultures. Laboratory of Psychometric and Evaluative Research Report.
INSTITUTION	Massachusetts Univ., Amherst. School of Education.
REPORT NO	LR-304
PUB DATE	2000-01-00
NOTE	30p.; Cover page varies.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Cross Cultural Studies; *Cultural Awareness; *Quality of Life; *Test Construction; Test Format; Test Items; *Translation; *Validity

ABSTRACT

Whatever the purpose of test adaptation, questions arise concerning the validity of inferences from such adapted tests. This paper considers several advantages and disadvantages of adapting tests from one language and culture to another. The paper also reviews several sources of error or invalidity associated with adapting tests and suggests ways to reduce those errors. It also considers test adaptation advances in a rapidly emerging area of social research, quality of life measures. The term "test adaptation" is preferred to test "translation" because adaptation is a broader term that better reflects what should happen in preparing a test that is constructed in one language and culture for use in another language and culture. Adapting a test may be cheaper than developing a new test, and it may allow better cross-cultural comparisons, provide a greater sense of security, and enhance fairness. Sometimes, test adaptation is not the answer, and when it is a good approach care must be taken to ensure validity. Sources of error that arise in test adaptation can be organized into three broad categories: (1) cultural and language differences; (2) technical methods; and (3) interpretation of results. Each of these must be considered in the adaptation process. The use of quality of life (QOL) tests by medical and public health researchers is a growing area that requires more than mere translation of the test. Suggestions are made to ensure that the adaptation of QOL tests receives the rigorous judgmental analysis it deserves. (Contains 28 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 207

Adapting Tests for Use in Multiple Languages and Cultures

Ronald K. Hambleton and Liane Patsula
University of Massachusetts at Amherst

Abstract

There is a growing interest in using tests constructed and validated for use in one language and culture in other languages and cultures. Sometimes these tests when adapted for use in a second language and culture can further research and meet informational needs, and other times, cross-cultural comparative studies can be carried out. But, whatever the purpose for the test adaptations, questions arise concerning the validity of inferences from these adapted tests.

The purposes of this paper are (1) to consider several advantages and disadvantages of adapting tests from one language and culture to another, (2) to review several sources of error or invalidity associated with adapting tests and to suggest ways to reduce those errors, and (3) to consider test adaptation advances in one rapidly emerging area of social research--quality of life measures.

TM033506

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Hambleton
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

BEST COPY AVAILABLE

Adapting Tests for Use in Multiple Languages and Cultures[1,2]

Ronald K. Hambleton and Liane Patsula
University of Massachusetts at Amherst

There is considerable evidence indicating that the need for multi-language versions of achievement, aptitude, and personality tests is growing. For example, the International Association for the Evaluation of Educational Achievement recently conducted the Third International Mathematics and Science Study in 45 countries that involved preparing mathematics and science tests in over 30 languages. This is the largest international study of its kind. A follow-up study is also being planned.

Prominent examples of new test adaptation projects in the United States include studies to prepare Spanish versions of College Board's Scholastic Assessment Test (SAT), American Council on Education's General Educational Development Tests (GED), and the United States Department of Education's National Assessment of Educational Progress (NAEP). All of these tests are immensely important in the United States--the SAT is used in college admissions, the GED as a high-school equivalency exam, and the NAEP for monitoring the quality of American education, and for the first time, these tests are being adapted into Spanish. Substantially more test adaptations can be expected in the future as (1) international exchanges of tests become more

---

[1]Laboratory of Psychometric and Evaluative Research Report No. 304. Amherst, MA: University of Massachusetts, School of Education.

[2]To appear in Social Indicators Research.

common, (2) credentialing exams are adapted into multiple languages, and (3) interest in cross-cultural research grows.

This need to translate or adapt psychological tests will be well-known to psychologists working in Europe and many parts of the world, but psychologists and other social science researchers in the United States are considerably less familiar with the need to adapt tests, and they are generally unfamiliar with methods for properly adapting tests and establishing their equivalence in a second language or culture (Geisinger, 1994; Hambleton, 1993; Hui & Triandis, 1985). Many American psychologists would be surprised, for example, to learn that Professor Charles Spielberger's popular measure of state-trait anxiety has been adapted for use in more than 50 languages. Also, popular American intelligence and personality tests are being widely adapted for use around the world (see, for example, Naughton & Wiklund, 1993).

The process of adapting psychological tests is all too often viewed as a simple task of finding someone who knows the languages, and having that person spend "a couple of hours" in getting the translation done. All too often, little concern is shown for any cultural differences which may exist and need to be addressed in a test adaptation and in the interpretation of results. One frequent problem centers on the presentation format of the test which may be less familiar to persons in one culture than another. The multiple-choice format, for example, is very familiar in the United States but substantially less familiar in other parts of the world.

The purposes of this paper are (1) to consider several advantages and disadvantages of adapting tests from one language and culture to another, (2) to review several sources of error or invalidity associated with adapting tests and to suggest ways to reduce those errors, and (3) to consider test adaptation advances in one rapidly emerging area of social research--quality of life measures. For our purposes in this paper, all types of paper and pencil instrumentation used by social researchers such as those which measure achievement, aptitude, personality, attitudes, opinions, and preferences will be referred to as tests.

In this paper, focus will center on test adaptation not test translation. The term test adaptation is preferred to the more popular and frequently used term test translation in our work because the term test adaptation is broader and more reflective of what should happen in practice when preparing a test that is constructed in one language and culture for use in a second language and culture. Test adaptation includes such activities as (1) deciding whether or not a test can measure the same construct in a different language and culture, (2) selecting translators, (3) deciding on appropriate accommodations to be made in preparing a test for use in a second language, and (4) adapting the test and checking its equivalence in the adapted form. Test translation, on the other hand, is only one of the steps in the process of test adaptation and even at this step, adaptation is often a more suitable term than translation to describe the actual process that takes place.

Construct a New Test or Adapt an Existing Test?

One of the first questions which arises in cross-cultural research is: Should a new test be constructed or should an existing test be adapted? There are a number of reasons for adapting a test. One reason is that it is usually cheaper and faster than preparing a new test for a second language group and may permit the use of existing test score norms (Geisinger, 1994). The development, validation, and norming of an individually administered intelligence test or personality measure can take several years and require substantial amounts of money for activities such as field testing, and compiling technical information and norms. In addition, by adapting an existing test, there is a database that provides a basis for designing and interpreting validity studies on the adapted test. Similarity of research findings between the original and adapted tests strengthens evidence for the validity of the test in its adapted form.

A second reason is that cross-national, cross-language, and/or cross-ethnic comparative studies require adapted tests. Such studies have become popular in recent years as many countries strive to set world-class educational standards and want to compare their progress to other countries. The Third International Mathematics and Science Study is a good example. Tests which are prepared in each country, even if the same test specifications are used, will be sufficiently different that valid comparisons of results across countries cannot usually be made.

A third reason is that researchers often have a sense of security which comes from adapting a test rather than initiating a new test development project in a second language and/or culture. However, some caution must be shown since even in the case of very frequently used tests such as the Minnesota Multiphasic Personality Inventory and the Internal-External Locus of Control scale, criticism may be in order, or improvements could be made in the adapted version of the test.

Finally, tests are sometimes adapted to enhance fairness by enabling persons to take tests in their preferred languages. For example, high school students in Israel can take their college admission exams in one of six languages. The bias in exam scores associated with students being forced to take their exams in their second or third best language is removed, and exam score validity is enhanced.

On the other hand, there are definitely times when test adaptation may not be justified. First, test adaptation is sometimes done because of the hope or mistaken belief that a valid test will result. But, it is not enough to depend on the popularity of the test in the original language. A researcher using an adapted test still has the responsibility of producing evidence of validity in the context where that adapted test is used. And, sometimes the construction of a new test may actually increase validity over the validity level achieved with an adapted test because a newly constructed test can be matched closely to the intended purpose of the test. With respect to personality assessment, according to Doug Jackson, one of the

leading researchers in the field of personality assessment, "Many of the most popular measures of personality were developed in an earlier era when our understanding of personality measurement was in its infancy and conceptual, quantitative and technological support for test construction was relatively primitive" (Jackson, 1991). Test adaptation would seem to have limited value for many of the older personality assessment tests.

A second disadvantage is that researchers risk imposing conclusions based on concepts which exist in their own cultures but which are foreign, or at least partially incorrect, when used in another culture. Tests provide operational definitions of certain concepts, however, there is no guarantee that those concepts, or those same operational definitions, exist in other cultures. A researcher must determine if the constructs measured in the original form are consistent with the goals of assessment after translation. This is a matter of assessing equivalence in construct relevance and construct operationalization (Helms, 1992; Hui & Triandis, 1985).

In practice, the decision to adapt a test versus construct a new test will depend on many factors--the nature of the study in which the test will be used (some purposes will require adapted tests); the time, resources, and expertise available; the relevance of the construct measured by the test across language and cultural groups; and the advantages associated with using an established test, for example, the availability of norms and a research base. In the next section, some of the major sources of

errors which arise in adapting a test and how they might be addressed will be considered.

Sources of Invalidity or Errors in Adapting Tests

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1985) provides direction for psychologists who select, develop, administer, and use educational and psychological tests. Three of the standards seem to be especially relevant in the context of test adaptation and use. Standard 6.2 states the need to revalidate a test when major revisions are made to it such as when a test is adapted for use in a second language. Standard 13.4 states the need to assess reliability and validity of adapted tests for their intended uses. Finally, Standard 13.4 states the need to establish the comparability of multi-language tests when comparability of tests is important.

These three standards provide a framework for considering sources of error or invalidity which might arise in test adaptation projects. These same sources of error and how they might be addressed in practice are reflected in the 22 International Test Commission guidelines for adapting tests from one language and culture to another (Hambleton, 1994; van de Vijver & Hambleton, 1996).

For our purposes, sources of error that arise in test adaptation can be organized into three broad categories: (1)

cultural/language differences, (2) technical methods, and (3) interpretation of results. Failure to attend to the sources of error in each of these categories can result in an adapted test which is not equivalent in the two language and cultural groups. Non-equivalent forms of a test, when the assumption is that the forms are equivalent, can only lead to errors in interpretation and faulty conclusions.

A good example of the misinterpretations which can follow from poor test adaptation is the following: In a recent international comparative study of reading, American students were asked to consider pairs of words and identify them as similar or different in meaning. "Pessimistic--Sanguine" was one of the pairs of words where American student performance was only slightly above chance. Only 54% of the American students answered the question correctly. In the country ranked first in performance, about 98% of the students answered the question correctly! In the process of attempting to better understand the reason for the huge difference in performance it was discovered that the word "sanguine" had no equivalent word in the language of this top performing country and so the foreign language equivalent of the English word "optimistic" was chosen. This substitution made the question considerably easier. In fact, pessimistic and optimistic are clearly words with opposite meaning, and would have been answered as such by a high percentage of the American students had they been presented with the pair of words "pessimistic--optimistic." The point of this example is to highlight the danger in drawing conclusions from

international studies without strong evidence that the test adaptation process resulted in two equivalent forms of the test. A careful review of the two words by translators should have detected the problem; a thorough analysis of empirical evidence collected from a field-test administration of the two versions of the test item would definitely have detected the improper translation.

Another example of faulty interpretations was provided by Charles Spielberger from the University of South Florida (personal communication). He noted that in some recent research, Japanese adults were found to be more depressed than their American counterparts. But when the topic was studied in more depth, a different explanation for the results was generated. Apparently, Japanese persons are more inhibited about admitting to positive feelings in their lives (as represented in statements in a personality test such as "I feel happy") than Americans. The consequence is that the Japanese score lower on psychological scales of wellness which make them appear more depressed when compared to Americans. Clearly, a false conclusion is drawn from the findings because of the failure to fully understand cultural differences between Japanese and Americans.

Each of the three categories of errors will be considered next.

1. Cultural/Language Differences Affecting Scores

The assessment and interpretation of cross-cultural results should not be viewed narrowly with the focus only on the equivalence of the words in the source and adapted versions of

the test. Rather, this process should be considered for all parts of the assessment process, including (1) construct equivalence (Does the test measure the same construct in each language version?), (2) test administration (Was each language version of the test administered in an identical fashion?), (3) test format (Will the format of the test be equally appropriate in each language version?), (4) speed of response (Will speed of response be more of a factor in one language version of the test than another?), and (5) other response styles such as acquiescence, tendency to guess, and social desirability.

One of the important factors in this category, Test Format, will be considered next.

Test Format. Differential familiarity with particular item formats presents an important source of invalidity of test results in cross-cultural studies. In the United States, selected response questions, such as multiple-choice questions, have been used extensively in assessment. In cross-cultural studies, it cannot be assumed that everyone is as familiar with multiple choice items as American students. Nationalities that follow the British system of education place greater emphasis on essays and short answer questions, as opposed to multiple-choice items. Thus, students from these countries are placed at a possible disadvantage as compared to their American counterparts. When constructed response formats such as essay questions are emphasized or serve as the dominant mode of assessment, persons with more experience with selected response formats such as multiple-choice items will be placed at a disadvantage.

Sometimes a balance of item formats may be the best solution to insure fairness and reduce sources of invalidity in the assessment process.

Another solution to the potentially biasing effect associated with a particular item format, is to include only those formats which groups being assessed have experienced. Whenever it can be demonstrated that respondents are not placed at a disadvantage, and when all variables of interest can still be measured, multiple-choice items or simple rating scales should be preferred because of the ease of data processing. But, even rating scales can be problematic. In one study, it was found that respondents had difficulty associating category descriptions such as "not at all" with "0" and "sometimes" with "1", etc. (Canales, Ganz, & Ciscarelli, 1995).

2. Technical Designs and Methods

A second category of errors is in the area of technical designs and methods. There are five main sources of errors that can influence the validity of adapted tests: (1) the test itself, (2) selection and training of translators, (3) the process of translation, (4) judgmental designs for adapting tests, and (5) empirical analyses for establishing equivalence.

With respect to (1), the test itself, when it is known in advance of test development that a test will be adapted into multiple languages and cultures, special precautions can be taken at the outset to maximize the suitability of the test in its multiple forms. Otherwise, problems in the selection of content,

format, etc. may need to be overcome at the test adaptation process.

With respect to (2), selection and training of translators, factors in addition to language proficiency are essential in a good translator--such as familiarity with the culture of the target group. Otherwise, an improper test adaptation may result.

With respect to (3), the process of translation, one concern centers on the presence of dialects and how they should be handled in test adaptation. Another concern is matching words from one language to the other based on frequency of use or familiarity. Failure to attend to these matters introduces error in the adapted test which can impact on test validity.

Concerning (4) judgmental designs for adapting tests, backward translation designs are popular but forward translation designs provide stronger evidence of test equivalence because both the source and target language versions of the test are scrutinized. That a test can be back-translated correctly (backward translation design) is not a guarantee of the validity of the target language version of the test. Unfortunately, backward translation designs are popular and yet fundamental errors are associated with this approach.

Finally, with respect to (5), empirical analyses for establishing equivalence, classical and modern analyses such as item analysis, factor analysis, structural equation modeling, and item bias detection (sometimes call "DIF" studies) (Holland & Wainer, 1993) are invaluable in detecting problems of non-equivalence of multi-language versions of a test. Unfortunately,

all too often, empirical analyses are not carried out, and errors in the test adaptation process go undetected.

One of these five categories of error, Selection and Training of Translators, will be considered in more detail next.

Selection and Training of Translators. The importance of obtaining the services of competent translators should be obvious. Too often though, researchers have tried to go through the translation process with a single translator selected because he/she happened to be available--a friend, a relative of a colleague, someone who could be hired very cheaply, etc. Competent translation work cannot be assumed from these sorts of persons. Also, the use of a single translator, regardless of the competency level, does not permit valuable interactions among independent translators to take place to resolve different points which arise in preparing a satisfactory test adaptation.

Translators, regardless of their numbers, should be more than persons familiar and competent with the languages involved in the translation. They should know the cultures very well, especially the target culture. Knowledge of the cultures involved especially the target culture is often essential for an effective adaptation. Also, subject matter knowledge in the adaptation of achievement tests is essential. The nuances and subtleties of a subject area will be lost on a translator unfamiliar with the subject matter. Too often, translators without technical knowledge will resort to literal translations which are often problematic to target language respondents and threaten test validity.

Finally, test translators will benefit from some training in test construction. For example, test translators need to know that when doing adaptations of achievement or aptitude tests they should not create "clang associations" that might lead test-wise respondents to the correct answers, or translate distractors in multiple-choice items unknowingly so that they have the same meaning. A test translator without knowledge of the principles of test and scale construction could easily make test material more or less difficult unknowingly, and correspondingly, lower the validity of the test in the target population.

3. Factors Affecting Interpretation of Results

A third category of errors arises at the interpretation of results stage. In large-scale cross-cultural studies, the purpose of the test is to provide a basis for making comparisons between various cultural/language groups, so as to understand the differences and similarities that exist. Sometimes cognitive variables are of interest and other times the focus may be on the assessment of personality variables or general information. Results should be used for seeking ways of comparing groups and understanding the differences. Cross-cultural studies should not be used to support arguments about the superiority or exceptionality of nations as if the international comparative study is the equivalent of a horse race with winners and losers.

In this context, to gain a better understanding when interpreting scores, other relevant factors external to the tests or assessment measures and specific to a nationality should also be considered to minimize errors in interpreting the results.

Curricula, educational policies and standards, wealth, standard of living, cultural values, motivation to take the test, etc., may be essential for properly interpreting scores across cultural/ language and/or national groups.

Quality of Life Tests and Approaches for Adapting Them

As is evident in the health-related literature of the past decade, the use of quality of life (QOL) tests by medical and public health researchers to measure the impact of medical intervention in clinical trials and to assess the outcomes of health care services has become increasingly important (Guillemin, Bombardier & Beaton, 1993). Spilker, Simpson, and Tilson (1992) conducted a review of the literature in 1991 and found that over 160 QOL tests were in use. More tests have certainly been developed since that time. More importantly for our purposes, there has been a major effort to adapt many of these QOL tests for use in multiple languages and cultures.

Although the majority of the QOL tests found by Spilker et al. (1992) were developed and intended for use in English-speaking countries, interest in QOL has definitely not been restricted to English speaking populations. On the contrary, countries have become more culturally diverse through immigration and as interest in cross-cultural comparisons of QOL grows, there is a great need for QOL tests to be made available in multiple languages. But the process of test adaptation has not been easy. Two typical problems follow.

First, sometimes the construct of interest may not even be suitable or meaningful in a second culture. As an example, consider the attention given to women and childbirth in different cultures. In some cultures, it is the norm for women to give birth in a hospital and to remain in the hospital for a minimum of one day and to rest further when they return home. In other cultures, it is the norm for women to deliver their babies at home and to resume their work a few hours later. It would not be meaningful to measure the quality of health care services provided to pregnant women in cultures where it is nonexistent. Furthermore, it would not be meaningful to compare such different cultures on their quality of health care services provided to pregnant women since it is nonexistent in some cultures.

Second, all too often a literal translation of an English QOL test into another language is prepared but this does not ensure that the test has the same meaning across languages and/or cultures (Guillemin et al., 1993). A simple example of this would be the following question evaluating the average birth weight of children: "Please check the birth weight of your child: ____0-2 lbs ____2.1-4 lbs ____4.1-6 lbs ____6.1-8 lbs ____over 8 lbs." To use a test that contained this question in France, for example, would require it to be not only translated into French, but also adapted according to the metric system which is used in France.

A more systematic approach to adapting QOL tests that goes beyond mere translation is needed. Such an approach should involve not only the literal test translation, but also the

adaptation of the test to take into account the different ways QOL health issues may be expressed in different languages and/or cultures (see, for example, Bullinger, Anderson, Cella, & Aaronson, 1993).

It seems clear that guidelines are needed to aid researchers in adapting QOL tests from one language and/or culture for use in another language and/or culture. At the same time, there are two main reasons for adapting QOL tests from one language and/or culture to another and these reasons need to be considered when developing guidelines for adapting QOL tests. One reason is simply to measure some important outcome in the context of a second language and/or culture group with no interest in cross-cultural or cross national comparisons. For example, a researcher may want to study general anxiety levels of Turkish persons and may have no interest in comparing Turkish persons to Americans on the basis of their general anxiety scores. If a test to measure general anxiety were not available in Turkey, the researcher would need to develop a new anxiety test or adapt an existing anxiety test. If the researcher chose to adapt an existing test, it would not be necessary that the original and adapted tests be strictly equivalent.

A second reason for translating and/or adapting QOL tests from one language and/or culture to another is to make cross-cultural comparisons. Steps need to be taken to ensure that all language versions of the test are equivalent so that fair and valid comparisons can be made between the people from different cultures.

Guillemin et al. (1993) proposed a set of guidelines to preserve equivalence in cross-cultural adaptation of health-related QOL tests. More recently, another group of researchers presented a set of steps for adapting a cancer therapy QOL test to compare different cultures on their cancer therapy (Bonomi et al., 1996). The steps and guidelines from these two studies can be combined to yield the following:

1. Translate the test to the target language (forward-translation).
2. Using the translated test in the target language, translate the test back to the original language (back-translation).
3. Review the original and adapted versions of the test, resolve any discrepancies between the two versions, and examine cross-cultural equivalence of the original and adapted versions. This step includes examining the equivalence in the meaning of words (semantic equivalence), idioms and colloquialisms (idiomatic equivalence), as well as in the situation evoked or depicted (experiential equivalence) and the concept explored (conceptual equivalence).
4. Pre-test the two versions to check for equivalence in original and adapted versions using judgmental procedures.

These four steps are a mixture of the popular forward and backward judgmental designs. On the other hand, they fall short of the ideal test adaptation process because no emphasis is given

to the compilation of empirical evidence to support the validity of the adapted test. How do respondents react to the test in each language version? Are these reactions equivalent? These are important questions to address prior to conducting studies on the construct of interest in multiple language or cultural groups.

In cases where there is interest only in adapting a test for use in another culture and cross-cultural comparisons are of no interest, these guidelines focusing on judgmental reviews may be sufficient. However, if the intent is to conduct cross-cultural comparative studies, more steps are needed to ensure cross-cultural equivalence.

The argument for the compilation of empirical evidence is the same one that is used in developing achievement, aptitude, and personality tests. Field-tests are always recommended because problems are usually detected that go unnoticed by reviewers in some type of judgmental process. The same argument applies to the use of a test in multiple languages and cultures.

The current state of many cross-cultural adaptations of QOL tests reflect the four steps above. Public health and medical researchers are to be commended on their use of judgmental procedures in adaptation, but more emphasis on the more difficult and perhaps costly steps of field testing is needed.

Our review of the literature showed that some QOL researchers will sometimes use empirical procedures to assess the reliability and validity of the adapted tests, however they

rarely compiled empirical evidence in advance of using the test in their research projects. In most cases, researchers report evidence of both the reliability and validity of QOL scores (Gregoire, de Laval, Mesters, & Czarka, 1994; Hutter & Wurtemberger, 1997; Wagner, Patrick, McKenna & Froese, 1996; Wild, Patrick, Johnson, Berzon, & Wald, 1995) in its adapted form and in the desired populations. Evidence of reliability usually involves a measure of internal consistency (such as Cronbach's coefficient alpha). Validity evidence, when it is compiled in a comprehensive way, includes a combination of content, construct, discriminant, and convergent validity through the gathering of population norms (Aaronson, Acquadro, Alonso & Apolone, 1992), the use of factor analysis and multidimensional scaling (Tuchler, Hofmann, Bernhart & Brugiatelli, 1992), and prediction studies (Baker, Jodrey, Zabora, Douglas, and Fernandez-Kelly, 1996).

Additional quantitative evidence to analyze the psychometric equivalence of tests prior to their use is necessary to improve the current adaptation process. Even among medical and public health researchers there seems to be consensus that there is a lack of psychometric analyses providing evidence that the different language or cultural versions of QOL measures are equivalent (Anderson, Aaronson & Wilkin, 1993; Krebs & Schuessler, 1989). Moreover, these researchers seem to agree that there is a need for guidelines to address such psychometric issues (Bullinger, et al., 1993).

Bullinger et al. (1993) proposed both a minimum and an optimum set of criteria for conducting test adaptation studies. At a minimum level, forward and backward translation studies, reliability and validity studies in each language and cultural groups on samples of at least 100, and clear documentation of the test adaptation process and findings, would be needed. At an optimum level, more translators would be used and more reviews of the translations would be conducted, and expanded efforts to establish empirically the equivalence of the test in multiple languages and cultures would be carried out. The steps described below, address the optimum criteria sketched out by Bollinger et al. (1993).

Additional steps could be included to improve the adaptation process of QOL tests. Such steps would include an initial step regarding the meaningfulness of comparing different cultures on the contents of the QOL test (Helms, 1992). Is it meaningful to compare the cultures on the content covered by the test? Is this content meaningful in all cultures of interest? For example, it has been found that acculturation can be a factor in responses that will influence the validity of any cross-cultural comparisons. If the comparisons are not meaningful, then it is not worthwhile to adapt the test. This step was not always considered in the QOL literature we reviewed. Of course, it may be in practice that only part of a test can be adapted, and it is on results from this part that cross-cultural comparisons can be made.

Other examples may be instructive. Content bearing on sports and exercise are often a component of quality of life tests in the United States, but they are rarely included in QOL tests from other countries. Topics that are fairly specific to a country or economic development may also have limited value in QOL tests for cross-cultural comparisons. For example, items such as "distance to a hospital" or "cleanliness" may have limited importance in Western QOL tests because these are givens in Western societies. They may be highly salient in QOL tests in Third World countries (Fons van de Vijver, personal communication).

A second step that was missing from many studies we reviewed was the compilation of evidence, as might be provided by a factor analysis about the factorial structure of the test in each language version (see, for example, Reise, Widaman, & Pugh, 1993). Do the same factors emerge? Is the internal structure of the test the same in each language version? These are important questions in establishing test equivalence across language and cultural groups (though not sufficient, according to Helms, 1992).

Finally, steps that address the equivalence of items from different language versions of a test need to be introduced into the QOL adaptation process. Steps to address the item equivalence of different versions of a QOL measure were not found in the literature of adapting QOL measures for use in different languages and/or cultures. Steps to assess the item equivalence

of different versions of a QOL measure include choosing a linking design to place scores from different version of a test on a common scale and conducting item bias studies (see, for example, Holland & Wainer, 1993).

There is a large number of statistical procedures to analyze item equivalence of adapted test items. Logistic regression, the Mantel-Haenszel procedure, and item response theory are three of the better known and useful approaches to identifying item bias due to cultural differences or flawed adaptations. The basic approach is to match respondents on the trait being measured by the test, and then to compare their performance on each item. When performance differs, questions about item non-equivalence are raised. Possibly the non-equivalence is due to cultural differences, or possibly the non-equivalence is due to a flawed translation. A major advantage of these approaches is that the findings are independent of any group differences which may exist. It is said that these approaches are "sample independent" (Hambleton, Clauser, Mazor, & Jones, 1993). For a comparison of several promising statistical approaches for detecting problems at the item level, readers are referred to a study by Budgell, Raju, and Quartetti (1995).

In summary, the cross-cultural adaptation of a test requires a rigorous judgmental analysis. The QOL research literature reflects evidence of the knowledge of these analyses and the steps can be found in the research literature. Unfortunately, to date, a judgmental analysis is often the only evidence offered in

support of test equivalence across multiple languages and cultures. In some instances, substantial evidence is compiled to address the reliability and validity of the adapted test. This is valuable and sufficient when there is no interest in cross-cultural comparisons. On the other hand, interest in cross-cultural comparisons requires rigorous evidence of test equivalence across language and cultural groups. The inclusion of steps in the test adaptation process to address (1) the meaningfulness of comparing different cultures on the content of the QOL test, (2) evidence to support the equivalence of the factorial structure of the test in multiple languages, and (3) compilation of empirical evidence addressing item equivalence, should provide the necessary evidence of equivalence to revise the test adaptation and/or to gain psychometric support for the equivalence of the test in two or more languages and cultures.

Conclusions

Interest has grown considerably in recent years in test adaptation. One of the main points of the paper is that researchers need to seriously consider whether an adapted test will serve the research better than developing a new test in the language and cultural groups of interest. They will also want to consider whether the construct being measured by the test is the same across language and cultural groups. Serious errors in interpretation can be minimized with concern for establishing construct equivalence first.

A second point of the paper is to focus researchers' attention on a wide variety of errors that arise in the practice of adapting tests, and to describe how these errors might be identified and minimized--for example, errors that result because of a failure to establish construct equivalence, errors that arise in the process of adapting tests such as those associated with the selection of item formats and test translators, and errors that arise in the interpretation of results.

Finally, a review was conducted of test adaptation initiatives in the area of quality of life tests. There is evidence of a growing methodology for the judgmental review of adapted tests, and concern for assessing reliability and validity of adapted tests in the intended populations. At the same time, with respect to tests intended for use in cross-cultural studies, there is an absence of methodology and studies that established the equivalence of QOL tests in multiple languages and cultures. More effort to establish the factorial equivalence and item equivalence of tests across language and cultural groups would seem to be in order because of the consequences on the validity of inferences from the test results.

References

Aaronson, N. K., Acquadro, C., Alonso, J., & Apolone, G. (1992). International Quality of Life Assessment (IQOLA) Project. Quality of Life Research, 1(5), 349-351.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Anderson, R. T., Aaronson, N. K., & Wilkin, D. (1993). Critical review of the international assessments of health-related quality of life. Quality of Life Research, 2(6), 369-395.

Baker, F., Jodrey, D., Zabora, J., Douglas, C., & Fernandez-Kelly, P. (1996). Empirically selected instruments for measuring quality-of-life dimensions in culturally diverse populations. Journal of the National Cancer Institute Monographs No. 20, 39-47.

Bonomi, A. E., Cella, D. F., Hahn, E. A., Bjordal, K., Sperner-Unterweger, B., Gangeri, L., Bergman, B., Willems-Groot, J., Hanquet, P., & Zittoun, R. (1996). Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Quality of Life Research, 5, 309-320.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. Applied Psychological Measurement, 19(4), 309-321.

Bullinger, M., Anderson, R., Cella, D., & Aaronson, N. (1993). Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Quality of Life Research, 2, 451-459.

Canales, S., Ganz, P. A., & Coscarelli, C. A. (1995). Translation and validation of a quality of life instrument for Hispanic American cancer patients: methodological considerations. Quality of Life Research, 4, 3-11.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6(4), 304-312.

Gregoire, J., de Laval, N., Mesters, P., & Czarka, M. (1994). Validation of the Quality of Life in Depression Scale in population of adult depressive patients aged 60 and above. Quality of Life Research, 3(1), 13-19.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. Journal of Clinical Epidemiology, 46, 1417-1432.

Guyatt, G. H. (1993). The philosophy of health-related quality of life translation. Quality of Life Research, 2(6), 461-465.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 54-65.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-244.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. European Journal of Psychological Assessment, 9(1), 1-18.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47(9), 1083-1101.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Publishers.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross cultural psychology. Journal of Cross-Cultural Psychology, 16, 131-152.

Hutter, B. O., & Wurtemberger, G. (1997). Reliability and validity of the German version of the Sickness Impact Profile in patients with chronic obstructive pulmonary disease. Psychology and Health, 12(2), 149-159.

Jackson, D. (1991). Problems in preparing personality tests and interest inventories for use in multiple cultures. Bulletin of the International Test Commission, 18, 88-93.

Krebs, D., & Schuessler, K. (1989). Life feeling scales for use in German and American samples. Social Indicators Research, 21(2), 113-131.

Naughton, M. J., & Wiklund, I. (1993). A critical review of dimension-specific measures of health-related quality of life in cross-cultural research. Quality of Life Research, 2, 397-432.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552-566.

Spilker, B., Simpson, R. L., & Tilson, H. H. (1992). Quality of life bibliography and indexes: 1991 update. Journal of Clinical Research Pharmacoepidemiology, 6, 205-266.

Tuchler, H., Hofmann, S., Bernhart, M., & Brugiatelli, M. (1992). A short multilingual quality of life questionnaire: practicability, reliability and interlingual homogeneity. Quality of Life Research, 1(2), 107-117.

van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1(2), 89-99.

Wagner, T. H., Patrick, D. L., McKenna, S. P., & Froese, P. S. (1996). Cross-cultural development of a quality of life measure for men with erection difficulties. Quality of Life Research, 5(4), 443-449.

Wild, D., Patrick, D., Johnson, E., Berzon, R., & Wald, A. (1995). Measuring health-related quality of life in persons with genital herpes. Quality of Life Research, 4(6), 532-539.